DOCUMENT RESUME

ED 181 029 TM 009 869

AUTHOR Gardner, Eric F.
TITLE Bias.
INSTITUTION National Council on Measurement in Education, East Lansing, Mich.
PUB DATE 78
NOTE 6p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978)
JOURNAL CIT NCME Measurement in Education; v9 n3 p 1-5 Sum 1978

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Bias; *Competitive Selection; Emotional Response; Personnel Selection; *Predictive Measurement; Selective Admission; Social Values; Standard Error of Measurement; *Statistical Analysis; Statistical Bias; Technical Reports; *Test Bias; Testing Problems

ABSTRACT

It is suggested that bias--when associated with a predictor, a test, or a statistical estimator--is not always bad, in spite of the immediate negative response evoked by the word, bias. Four settings are described to illustrate situations in which a procedure should not be summarily rejected due to bias: (1) educational researchers rejected the use of critical ratios for hypothesis testing, in favor of the t-test, to avoid bias which was negligible for most sample sizes; (2) ridge regression, which uses biased estimators, provides better prediction with smaller errors than do the less biased least squares estimators; (3) a utility model is proposed for competitive selection problems because it requires an explicit public statement of utilities (biases) for each subpopulation; and (4) low scores on an achievement test by a particular group do not necessarily prove that the test is biased against that group, when the test has appropriate content validity. It is concluded that bias should be recognized and explicitly considered when choosing the best solution from different alternatives, rather than pretending bias doesn't exist, fearing bias irrationally, or introducing other unknown biases in opposition. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED181029

Volume 9, No. 3

NCME measurement in education

Summer, 1978

A SERIES OF SPECIAL REPORTS OF THE NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Herbert C. Rudman
NCME

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Bias*

Dr. Eric F. Gardner
Syracuse University

ABOUT THIS ISSUE

In this era of equal opportunity, procedural safeguards, and nondiscriminatory testing procedures, the word *bias* carries so many negative connotations as to appear inappropriate as a title for an ME publication. In his presidential address to NCME, Eric Gardner discusses bias in a most enlightening fashion, injecting a note of hesitancy into the emotional overreaction associated with the term.

Dr. Gardner needs no introduction. He has spent the last 25 years documenting his pursuit of excellence. His accomplishments include numerous professional publications and presentations as well as many tests and test manuals, most significantly the Stanford Achievement Tests, the Stanford Diagnostic Reading Test, and the Stanford Diagnostic Mathematics Test.

He is currently the Margaret O. Slocum, Professor of Psychology and Education, Syracuse University, Syracuse, NY 13210.

PSR

TM009 869

*Presidential address presented at the annual meeting of the National Council on Measurement in Education, Toronto, March, 1978

Introduction

I'm sure that after noting the title of my remarks and being knowledgeable about the popular and professional outcries against bias, and the extensive research on materials containing various kinds of bias, ranging from test bias, selection bias, biased predictors, and just plain personal bias, you are wondering what I could possibly say that would make a contribution to the literature on bias. I do have a definite purpose even though it is a modest one. My thesis is that bias when associated with a predictor, test, or statistical estimator isn't always bad even though it inevitably evokes an immediate negative, *irrational* reaction. In fact, we, including the individuals we assumed are being harmed by bias, may lose unnecessarily when the use of a product or process associated with the adjective *bias* is summarily rejected for that reason alone.

I am going to make my point by using illustrations from three very different settings. The first setting requires a consideration which goes counter to one of the most sacred tenets in the pristine area of theoretical statistics. It requires a judgment which might lead to the choice of a definitely biased estimator over an unbiased estimator of the same parameter — *an unthinkable crime*. The second involves a consideration of bias in the selection for admission to educational institutions or for employment. The third is in the more general, more commonly encountered area of test bias.

Biased Statistics

Some of you are old enough to remember when a test of hypotheses about means was accomplished with the use of a critical ratio. For the benefit of the non-oldtimers, this was obtained by dividing the deviation from the hypothesized mean by the standard deviation of the sample (computed by $\sqrt{\sum \frac{x^2}{n}}$) and using the normal distribution as the appropriate sampling distribution. More of you can remember when the mark of a sophisticated educational researcher was one who castigated the users of critical ratios, and who insisted that the only possible appropriate

statistic was the t-test with its exact sampling distribution and unbiased estimate of the population variance. At that time there were no voices raised to point out that the bias in the variance was n-1, which for a large sample, say 100 or over, introduced negligible bias, and that for samples over 30 the difference in the tail areas obtained from segments of the t-distribution and normal distribution were almost identical. My point is merely that in the name of *bias* there was much ado about nothing, especially considering the size of most samples we use in measurement research. The situation would have been less ludicrous if the same people had not been so completely, and erroneously happy that the problem of bias had been solved for standard deviations by taking the square root of the unbiased estimate of variance. In any event, whatever bias existed was very minor, and neither great harm nor gain resulted from a consideration, or lack of consideration of bias in estimates of standard deviation and variance for most research endeavors.

My second example, however, is an attempt to make a different point about a biased statistic. For years we have been taught and have taught others that the best possible regression equation results when we set up a linear function and minimize the sum of the squares of the errors of estimate. Such a regression equation with its unbiased estimates of population regression coefficients, gives us the best (in the sense that the multiple correlation coefficient is largest and the standard error of estimate smallest) linear combination of predictors. What we have believed and taught, as many of you know, is correct as long as we accept only unbiased estimates (certainly a praiseworthy condition) or our statistics. If, however, we are willing to consider a trade-off between bias and reduction in mean squared error, the situation is substantially different. Relinquishing the requirement that the regression weights must be unbiased estimates leads to a procedure called ridge regression (Hoerl & Kennard, 1970).

Ridge regression, which uses biased estimates, can in many instances, give better prediction with smaller errors of estimate than could be obtained with the classical regression approach. This apparent paradox is due to the possibility of a substantial reduction in the size of the standard error which, when added to a small error due to bias, gives a much smaller *overall error* than the standard error of the statistic under the classical assumptions of unbiased estimates. Here is an illustration where the acceptance of known bias in a statistical estimate results in a much smaller error than can be obtained by the insistance upon having only estimates without bias. Mathematical statisticians have recently been searching for a variety of estimators (both biased and unbiased) which are more precise than least-squares estimators. (Andrew, Bickel, Hampel, Huber, Rogers & Tukey, 1972.)

Bias in Selection

The problem of bias in selection, whether for admission to educational institutions or for employment, has been, and still is, a troublesome one for a society in which equality of opportunity has been a major accepted philosophy. A number of approaches have been proposed as solutions to this problem ranging from a least squares regression method to various types of procedures based on specific ratios and probabilities. In a recent volume of *Journal of Educational Measurement* (*13,* 1976) devoted entirely to bias in selection, Peterson and Novik analyzed and evaluated the most widely used models and proposed one under the assumption that there exists an available, relevant, reliable criterion score, and an unbiased measure of performance for applicants in each subpopulation. The criterion problem, as always, is a difficult problem in itself. In any event, they proposed a threshold utility model which selects those applicants with highest expected utility. The utilities will differ for different subpopulations and the prediction equations may also differ from subpopulation to subpopulation.

NCME

measurement in education

Volume 9, No. 3 Summer, 1978

Paul S. Redelheim, Editor

Advisory Committee:
Sidney E. Benton, North Georgia College
Carmen Finley, American Institutes for Research
George S. Ingebo, Portland Public Schools
Anthony J. Polemeni, New York City Board of Education
Warren W. Willingham, Educational Testing Service

Measurement in Education is a series of special reports published four times a year in October, January, March, and May by the National Council on Measurement in Education. These reports are concerned with the practical implications of measurement and related research and their applications to educational problems of individuals, institutions, and systems. The emphasis is upon uses of measurement rather than technical or theoretical issues. Subscription rate: $5.00 per year; single copies 75¢ each in quantities of 25 or more, or $1.50 for a single issue.

Second-class postage paid at East Lansing, MI 48823 and at Cleveland, OH

Address Editorial correspondence to:

Dr. Paul S. Redelheim
George Peabody College for Teachers
Box 158
Nashville, TN 37203

Address Business correspondence to:

Dr. Irvin J. Lehmann
202 S. Kedzie Hall
Michigan State University
East Lansing, MI 48824

Their criticism of some of the models they evaluated is that these models are accepted solely because they yield a popular result: namely, they give lower cut scores for disadvantaged minority populations. As they say:

> The appeal of these models, then, is that they produce a desirable result . . . (However) one must be sure that the model is giving the right results for the right reasons. If the models are giving the right results for the wrong reasons, it may well be possible that, in some other circumstances, wrong answers will be forthcoming . . . . We shall show that these models sometimes can produce most undesirable results and could, in fact, be used to justify discrimination against some minority groups.

They very carefully point out that a major advantage of their model is that "it requires an explicit public statement of utilities for each subpopulation." The basic issue and problem is agreement on appropriate utilities for each subpopulation. Although their solution takes into account psychometric and statistical biases, it does not allow one to turn a crank and obtain an answer to the basic political, social, ethical and legal issues involved. In other words, a specified utility which is described as fair by one person or group is claimed to give biased results by another person or subpopulation. Thus bias in outcome, defined here as unfairness, is inevitable unless there is complete agreement on the appropriate utilities to use. Hence, as desirable psychometrically and statistically as the threshold utility model is, it does not permit a more satisfactory definition of "fair" selection than the statement from the latest (1974) APA/AERA/NCME *Standards for Educational and Psychological Tests*, namely:

> It is important to recognize that there are different definitions of fairness, and whether a given procedure is or is not fair may depend upon the definition accepted.

It does have the real virtue, as does Darlington's model, of forcing one to define bias or lack of bias in terms of utility for some social need. However, it does not remove the bias issue. It separates the political and social issues from the statistical. To quote Cronbach, "If discussion within the psychometric community helps the philosophers and lawyers to put more substantial arguments behind competing rules for attaining equity, we will have done our part."

Part of the difficulty with models such as the Constant Ratio or Conditional Probability Model (or, for that matter, the Regression of Equal Risk Models) is that they attempt to provide a purely technical resolution to an issue that involves value judgments. Values are implicit in many models. To adequately address the issues of values, they need to be dealt with explicitly. Decision theory provides a framework for doing this and that is the approach advocated by Peterson and Novick. *Note that it provides a framework for making a decision* but it doesn't automatically provide the decision or answer. The answer involves a value judgment which implies bias by somebody or bodies against A and for B and vice versa.

Test Bias

The literature on test bias and other comparable terms such as set and culture free tests is far more voluminous than that on bias in selection. The use of the word bias in this context has even more connotations than it does when associated with selection. Although the possibility of bias is present in all tests, I am going to restrict my comments to bias in achievement tests. In some instances a bias does exist, as I shall explore in the next section, and the bias may fail to be claimed or be claimed for the wrong reasons. But more often and what we are concerned about here is that when it doesn't exist at all, it may be claimed for irrelevant reasons. For example, as Bob Ebel pointed out in a previous issue of *Measurement in Education* (1977), some standardized educational achievement tests have been attacked for their alleged bias against cultural minorities and pupils of low socio-economic status because such pupils tend to make lower scores.

Lower scores *alone* on an achievement test do not signify bias. If they did, then every spelling test would be biased against poor spellers, every vocabulary test against persons who had poor vocabularies, and every shorthand test against persons who had never learned shorthand. The fact that a group of pupils who do poorly on a particular language test might do better if the test were in the Chicano dialect does not mean that the original test is biased against them. It means that they have poor English language skills, a situation that needs remediation. Because a pupil does poorly on a reading comprehension test but well on a comparable listening comprehension test does not mean the reading comprehension test is biased against him. The two contrasting scores give important information for instructional purposes. An achievement test is designed to measure the competence of pupils in specifically defined areas under conditions described in the manual. The score of each pupil indicates how adequately the questions defining the domain tested were answered. For a test to have satisfactory content validity, it is important that aspects of the variable the test was designed to measure *not* be deleted in the name of bias.

In contrast, a test designed to measure achievement in elementary science, and in which the items are presented in language and syntax far above the level of the pupils tested, *is* biased especially against the poor readers. A distinction must be made here between that aspect of reading difficulty based on science vocabulary which was part of the curriculum, and other vocabulary which was not. The score on the test should measure knowledge of science including appropriate scientific vocabulary, but the score should not depend upon general reading ability. Hence, the test maker should recognize the bias introduced by complex reading material and remove its confounding effect.

Time does not permit a consideration of even the most important variables in test bias such as motivation, negative reaction to specific items, set, etc. However, since cultural bias and sex bias are considered so important today, I would like to comment on each.

Some time ago when we were doing some preliminary work with very young children prior to developing a test of knowledge of science for the lower levels of the *Stanford Achievement Test*, Dick Madden obtained some interesting data. A person who was knowledgeable of the Chicano dialect converted (not translated literally) the written Stanford test into written Chicano. The two tests were then administered both orally and in written form to samples of Chicano pupils. The best measure of these pupils' knowledge of the scientific content we were trying to measure was a combination of their performance on the written English and oral Spanish. The poorest performance was on the written Spanish. When one considers that Spanish was spoken in most homes but

rarely read, the outcome does not seem so surprising. In this case — if our aim had been to measure the extent of their knowledge of science — the test written in Spanish would have been definitely biased against the Chicanos. We would have obtained a definite underestimate.

Recently it has come to my attention that the lowest age grade level of a diagnostic reading test has been rejected by a state committee because of sex bias. One of the major criticisms was that the items portrayed mostly traditional roles for males and females. The criticism was not made of the higher level tests. Here is an illustration where trade-offs are necessary. If the objective of the test was to promote the concept of sexual equality, and that sex is irrelevant to the roles in life one assumes, then it was clear that the test was not designed for that purpose. If the purpose was to present stimuli familiar to very young children just beginning to learn to read, then females in the roles of mother, teacher, etc., and males as father, mail carrier, and fireman are reasonable. The aim should be to avoid contaminating the reading task with children's experience or lack of it in their immediate environment.

Most very young children so far have not been exposed to very many males or females in "atypical" roles, e.g., *male* first grade teachers are not common. These children's experiences would have been relatively limited to what takes place at home or in school; and for the most part these are female dominated. This situation may change in time, but doesn't exist at present. It seems to me that efforts to promote a broader acceptance of sex roles can better be achieved in other parts of the curriculum rather than introducing this issue in a situation which would make the reading task more difficult. *Here is a value judgment.* The decision could be made either way, but should be done explicitly not because one attaches the label "sex bias" to a good diagnostic reading test.

Conclusion

In conclusion, I am making a plea for a cognitive consideration of the effect of a single important word, namely *BIAS*, which almost by definition evokes an affective, even visceral, reaction. The effect of bias in itself is undesirable (I wish I could think of a stronger word), but the elimination of everything to which the word bias could be related (including tests), or maintaining it doesn't exist, frequently results in less productive efforts to solve social issues where bias is of major concern.

I am requesting that you consider the adjective biased when associated with a noun, noun clause or phrase not as a NO-NO, but look on it as providing very valuable information; namely, that perhaps for one of the few times in life you have an opportunity to examine a situation where you have been given a specific warning signal. Who among us is not biased with respect to some variables or situations? Some of the bias is conscious and some unconscious. Some bias, e.g., bias in favor of virtue is good; bias in favor of evil or discrimination is bad. The fact that we have different backgrounds, experiences, aspirations, and values insures that a completely unbiased, objective individual doesn't exist. A situation which can be described as unbiased or fair for some individuals is considered biased or unfair by others.

REPORTS AVAILABLE

Back issues of *Measurement in Education* are available for 1.50 each per single issue, and for 75¢ each in quantities of 25 or more for a single issue.

Vol. 1,	No. 1	*Helping Teachers Use Tests* by Robert L. Thorndike
	No. 2	*Interpreting Achievement Profiles — Uses and Warnings* by Eric F. Gardner
	No. 3	*Mastery Learning and Mastery Testing* by Samuel T. Mayo
	No. 4	*On Reporting Test Results to Community Groups* by Alden W. Badal & Edwin P. Larsen
Vol. 2,	No. 1	*National Assessment Says* by Frank B. Womer
	No. 2	*The PLAN System for Individualizing Education* by John C. Flanagan
	No. 3	*Measurement Aspects of Performance Contracting* by Richard E. Schutz
	No. 4	*The History of Grading Practices* by Louise Witmer Cureton
Vol. 3,	No. 1	*Using Your Achievement Test Score Reports* by Edwin Gary Joselyn & Jack C. Merwin
	No. 2	*An Item Analysis Service for Teachers* by Willard G. Warrington
	No. 3	*On the Reliability of Ratings of Essay Examinations* by William E. Coffman
	No. 4	*Criterion-Referenced Testing in the Classroom* by Peter W. Airasian and George F. Madaus
Vol. 4,	No. 1	*Goals and Objectives in Planning and Evaluation: A Second Generation* by Victor W. Doherty and Walter E. Hathaway
	No. 2	*Career Maturity* by John O. Crites
	No. 3	*Assessing Educational Achievement in the Affective Domain* by Ralph W. Tyler
	No. 4	*The National Test-Equating Study in Reading (The Anchor Test Study)* by Richard M. Jaeger
Vol. 5,	No. 1	*The Tangled Web* by Fred F. Harcleroad
	No. 2	*A Moratorium? What Kind?* by William E. Coffman
	No. 3	*Evaluators, Educators, and the Publics: A Detente?* by William A. Mehrens
	No. 4	*Shall We Get Rid of Grades?* by Robert L. Ebel
Vol. 6,	No. 1	*On Evaluating a Project: Some Practical Suggestions* by John W. Wick
	No. 2	*Measuring Reading Achievement: A Case for Criterion — Referenced Testing and Accountability* by S. Jay Samuels and Glenace E. Edwall
	No. 3	*Security in a Citywide Testing Program* by Anthony J. Polemeni
	No. 4	*Dear Mama: Why don't they love me anymore?* by Thomas J. Fitzgibbon
Vol. 7,	No. 1	*The Discrepancy Evaluation Model (Part I)* by Andres Steinmetz
	No. 2	*The Discrepancy Evaluation Model (Part II)* by Andres Steinmetz
	No. 3	*What Graduate and Professional School Students Think About Admission Tests* by Letnard L. Baird
	No. 4	*The Paradox of Education Testing* by Robert L. Ebel
Vol. 8,	No. 1	*FACE UP: A Framework for Assessing Career Educators Utilization Programs* by Bruce W. Tuckman
	No. 2	*Grade Equivalent Scores* by Gary Echternacht and *If Not Grade Equivalent Scores — Then What?* by Jeanne M. Pias
	No. 3	*Criticisms of Testing: How Mean is the Median?* by Henry S. Dyer
	No. 4	*The Superintendent and Testing: Implications for the Curriculum* by Herbert C. Rudman
Vol. 9,	No. 1	*A Consumer's Guide to Criterion — Referenced Test Item Statistics* by Ronald C. Berk
	No. 2	*Perspective on Intelligence Testing*, by Roger T. Lennon

We have advertising and advertising slogans which offer assurance of objective research and research results, evaluation and evaluation results where the word or connotation *UNBIASED* or *OBJECTIVE* is presented either in ringing tones, or worse still is conveyed successfully with maximum subtlety. The fact is that explicit consideration of bias in a situation enables the people involved to consider various trade-offs among the various biases recognized and to make decisions which are overt and the best possible ones for all concerned. I suggest that we follow the directions we give for answering a multiple choice item, namely, "Choose the best answer from among those given" — in this case, the various possible alternatives which are related to the various biases. This direction implies that we look at all the alternatives and choose the one that is most appropriate for a particular problem. It precludes eliminating excellent possibilities, either because they are biased or because they appear biased to us. I have not lived long enough yet to see any decision which does not appear to be biased to someone for some purpose. We never eliminate bias completely but, by our attempts, merely transfer the effects of bias from one element, person, or situation to another.

I hope that no one present, either because of personal bias or my inability to be sufficiently articulate, will conclude that I am in favor of unfair practices, biased tests that produce invalid scores, or biased estimates that give erroneous pictures of their parameters. I am convinced that all activities by humans are, to some degree, idiosyncratic in that they involve personal points of view which reflect their environment, past experience and learned attitudes. Many of these unknown biases are not at the conscious level. Many, which are recognized by no one, including the person involved, may have a very *devastating* or sometimes very *desirable* effect on the person and his/her influence on others (society). It is far better to identify explicitly the various biases which can be ascertained and make appropriate decisions accordingly, than it is 1) to maintain that no bias exists, or 2) fail to use appropriate methods and tools because of an irrational fear of bias, or 3) in an attempt to remove a suspected bias, introduce *unknown* biases in opposition to one's basic objectives.

References

Andrew, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., & Tukey, J.W. *Robust estimators of location: survey and advances.* Princeton, N.J.: Princeton University Press. 1972.

APA/AERA/NCME *Standards for educational and psychological tests.* Washington, D.C.: Author, 1974.

Ebel, Robert. A paradox of educational testing. *Measurement in Education, 7(4),* 1977.

Hoerl, A.E., & Kennard, R.W. Ridge regression: Bias estimation for nonorthogonal problems. *Technometrics,* 1970, *9,* 55-68.

Peterson, N.S. & Novick, M.R. An evaluation of some models for culture-fair selection. *Journal of Educational Measurement,* 1976, *13,* 3-29